文章编号: 1006—9348(2005)05—0226—05

# 面向对话文本的自动摘要系统的研究

陈 卫平, 王永成, 刘传汉
(上海交通大学计算机科学与工程系, 上海 200030)

摘要: 该文介绍了对话文本自动摘要系统的一些关键技术, 包括体裁的识别、对话信息单元的识别、问题句与回答句的关联等。摘要的连贯性是衡量摘要质量的一个重要指标, 由于对话文本本身的交互性, 使得摘要的连贯性常存在于不同的对话者的对话内容之中, 并以问题—回答对的形式出现。该文设计了一种自动识别这些局部连贯性的方法, 该方法首先自动识别出所有的问题句, 然后识别出与问题句相对应的回答句, 形成问题—回答对; 最后根据启发式规则, 从这些问题—回答对中选取句子生成摘要。实验结果表明, 该方法具有较高的识别准确率, 并在无损摘要信息量的基础上大大提高对话文本摘要的连贯性。

关键词: 对话文本; 连贯性; 问答对; 语句相关度
中图分类号: TP181 文献标识码: B

# Research on Automatic Summarization of Spoken Dialogues

CHEN Wei—ping, WANG Yong—cheng, LIU Chuan—han
(Department of Computer Science and Engineering, SJTU, Shanghai 200030, China)

**ABSTRACT**: Automatic summarization of spoken dialogues is a relatively new area. Some critical techniques are proposed in this paper: (1) detection of spoken dialogues; (2) detection and linking of cross—speaker information units (question—answer pairs). Due to the interactive nature of dialogues, local regions of coherence often stretch across different speakers. An approach to automatically detect those regions of local coherence is presented. Firstly, all questions are detected. Secondly, all corresponding answers of each question are detected to constitute question—answer pairs. Lastly, some sentences are extracted from the question—answer pairs to comprise a complete summarization. Experimental results show that the approach is highly efficient and it will increase summary fluency significantly while not compromising informativeness.

**KEYWORDS**: Spoken dialogs; Fluency; Question—answer pairs; Sentence similarity

## 1 引言

文本的自动摘要发展到今天, 其方法主要分为两大类: 一类是基于理解的自动摘要方法, 另一类是基于统计的自动摘要方法[1,2]。

基于理解的自动摘要方法建立在人工智能、自然语言理解的基础之上, 利用语言学知识对文章进行复杂的语法分析、语义分析、语用分析, 最后生成摘要。由于这类方法实现了对文章主题内容的理解, 并且许多摘要的句子都是由系统自动生成的, 因此, 从质量的角度来看, 摘要反映原文主题的程度相对较高, 且语句精炼、连贯性好。我们知道, 目前人工智能与自然语言理解还存在着许多难以解决的问题, 使得目前这类方法还不能够得到快速的发展。另外, 这类方法与文章所涉及的领域关系非常密切, 从而使得基于这类方法的自动摘要系统的可移植性较差。

基金项目: 863 计划资助项目(2002AA119050)。
收稿日期: 2004—07—16

基于统计的自动摘要方法则是根据文章的篇章结构, 利用统计的方法来获取文章的主题内容。基于统计的自动摘要方法处理速度快, 对于一般结构比较规范的文章处理效果较好。由于这类方法过分依赖于文章结构的规范性, 且对句子或段落没有进行语义分析, 因此存在着明显的不足, 特别是对包含有多个主题的文章进行摘要时, 常常出现遗漏主题或者摘要不连贯等现象。

对话式体裁文本(简称对话文本)属于一种特殊的多主题文本, 在此类文本中一般会出现两个或两个以上的对话参与者。他们中的一方接受(或拒绝)另一方的请求, 提出(或回答)问题, 或者对其他对话参与者的言论进行确认(或评论)等等。如果将传统的基于非对话文本的自动摘要方法应用于对话文本, 摘要的连贯性往往会出现这样的问题, 即参与对话的一方与另一方的对话不匹配, 大大降低了摘要的连贯性、可读性和逻辑的正确性。随着语音对话数据库的大量出现, 以及诸如人物访谈、记者招待会、视频节目的网上文字直播等等数量巨大的网上对话文本的出现, 人们迫切需要一

个高质量的对话文本的自动摘要系统。尤其是，该系统将有助于诸如多方会议、买卖交易、客户洽谈等各种各样的对话记录的分类、标引和检索工作。

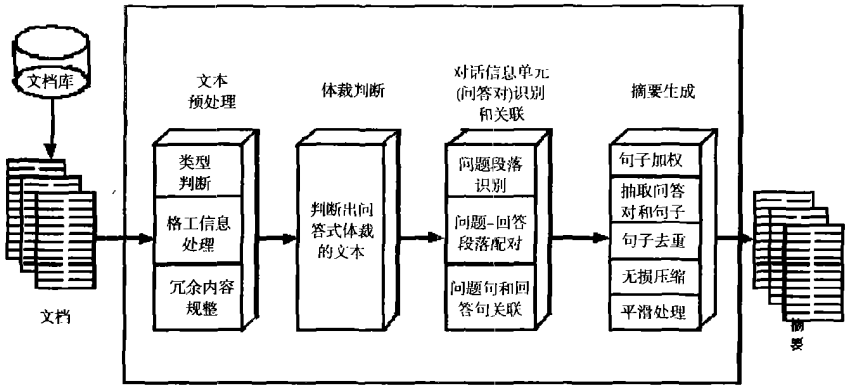


图 1　系统整体结构图

本文在基于统计的自动摘要方法的基础上，给出一种针对对话文本的自动摘要的新方法。本文将依次介绍对话文本的体裁判断，问题段落的自动识别，问题段落和相应回答段落的配对，以及问题句与回答句之间的关联，并给出一种计算问题句与回答句相关度的方法。在最后生成摘要时，就可以利用问题句和回答句之间的关联信息，避免传统方法针对对话文本进行摘要所产生的问题，大大提高摘要的整体连贯性和可读性。实验结果表明，针对对话文本，该方法具有较高的体裁识别率和比较满意的问答句配对的准确率，使得在不损失摘要信息量的基础上大大提高对话文本摘要的连贯性。

## 2　系统概述

基于对话文本的自动摘要系统（如图 1 所示）主要由以下四个部分组成：文本预处理、体裁判断、对话信息单元（问答对）的识别与关联、摘要的生成。

文本预处理主要是将不同类型、不同文件格式的文本进行处理，得到类型单一、格式规范、特征突出的纯文本。

体裁判断（或称体裁识别）就是判断输入的文本是否属于对话式体裁文本。一般来说，文章的体裁多种多样，专家对不同体裁的文章进行手工摘要时，其方法也都不尽相同，因此，针对某种体裁的文本，采用相适宜的自动摘要方法，将有利于提高自动摘要的质量。

对话信息单元（问答对）的识别与关联就是在正确识别了体裁以后，针对对话文本的特点，为了提高摘要的一致性，对文本中的问题句和相对应的回答句进行配对，简单的说，就是首先识别出每个问题句，然后再查找每个问题句所对应的回答句。

摘要的生成就是运用启发式规则，对文本中的句子及问答对进行加权和选取工作，并将选取的句子有机地组成摘要文本。

## 3　关键技术的设计和实现

### 3.1　体裁判断

我们首先进行对话式文本体裁的识别，其结果将决定下一步处理的方法是采用传统的方法还是采用我们设计的特定方法。我们经过对大量样本的统计分析，发现对话式体裁文本具有一般的特征，利用这些特征就可以判断出所给文本是否属于对话式体裁的文本。

我们从收集的 9562 篇对话式体裁文本（来自人民网、新华网、新浪网等网站）中随机抽取出 2000 篇，经过人工初步统计和深入分析，发现这类体裁的文章中所有参与的对话者信息都会在段落的首句出现，并以冒号或者其他符号作为结束标记。从表 1 的统计数据可以看出，99.6% 的对话文本都有明显的标记来标识对话者。利用这些标记的特征信息，统计出所有出现在段落首句，并以特征标识“：”、“:”、“】”、“]”等结尾，且满足一定长度限制的字串的个数，再结合文章的长度，我们就可以判断出该文本是否为对话文本。

表 1　对话者特征信息统计表

| 结束标记 | 样 例 | 篇数 |
|---|---|---|
| ：或: | 问: 最难过的是什么时候？ | 1968 |
| 】或] | 【问:】你怎么看待新闻自由？你自由吗？[网友六神无主]　徐老师，您觉得新东方的成功是否意味着传统正规大学的教育的失败呢？ | 24 |
| 其他 | 今天上午布什总统谈及台湾问题时是否使用了”一个中国政策”这个字眼？ | 8 |

### 3.2　对话信息单元（问答对）的识别和关联

所谓对话者信息单元（问答对）是指，将文章中出现的所有的对话部分的内容，以两个对话参与者的对话信息作为一个单位进行的划分。问答对的识别和关联的好坏将直接影响到后面摘要生成的质量，即摘要的连贯性、可读性以及准确性。我们首先识别出提问者的问题段落，然后找出相对应的回答者的回答段落，最后将每个具体的问题句与其相对应的回答句进行配对关联。在这里，我们进行问题回答段落的识别与配对的目的是为了进一步提高我们后面的问题句和回答句的关联正确率，并且减少语句之间相关度计算的工作量。

#### 3.2.1　问题段落识别

问题段落识别的主要任务就是识别出文章中出现的所有问题段落。问题段落识别分为以下几个步骤进行：统计出

所有参与对话者的信息；判断对话者的身份（即判断对话者是提问者还是回答者）；利用对话者身份信息识别出所有问题段落。

1）统计对话者信息

对话者信息包含三个方面的内容：对话者的标识字串；对话者标识字串在文中出现的次数；对话者的身份，其数据结构定义为：

```
struct speaker{
    sstring     speaker_tag;
    int         occurance_number;
    int         identity;
};
```

由于所有参与对话者的信息都出现在段落首句，所以可以通过对各个段落首句的从前至后的一次扫描，就可以按顺序统计出文中出现的所有对话者的标识字串（speaker_tag）信息和标识字串出现的次数（occurance_number）信息。

2）判断对话者的身份

我们运用统计的方法，判断某个对话者是提问者还是回答者。首先建立一个问题特征词典，其中包含问题特征词（根据对 2000 篇样本的统计分析，我们目前发现了常用的 100 多个标识问题的特征词，如：？，谁，为什么，何时，多少，等等）；然后运用问题特征词典，综合考虑这些问题特征词在当前对话者话语中出现的频次以及对话信息的长度等因素，判断这一段落是否为问题段落；最后统计出每个对话者提问问题的段落次数（即其作为提问者出现的次数），如果这个次数大于该对话者在文中出现次数的一半，就可以判定该对话者为提问者；反之，则为回答者（如图 2 所示）。

3）问题段落标识

借助对话者的身份信息，可以将身份为提问者的所有对话信息标识为问题段落。

3.2.2　问答段落的配对

据 Klaus Zechner 统计[4]，在对话文本中，每个问题都有至少一个回答，而且一般回答都在问题的后面，很少会出现连续两个问题段落而只有一个回答段落的情况。针对这个特点，可以将每两个问题段落之间的部分都作为前面一个问题段落的回答段落，最后一个问题段落后面的所有部分均作为最后一个问题段落的回答段落。利用这种简单的规则，我们便可以标识出与每个问题段落相对应的回答段落。

3.2.3　问题句和回答句的关联

由于每个问题段落中可能包含有多个问题句，其对应的回答段落中也就有相对应的很多回答句，因此，在识别出每个问题回答段落以后，再来考虑如何在一个问题回答段落中让每个问题句都与其对应的回答句进行配对。在这里，我们采用语句相关模型来计算问题句和回答句之间的相关度[5]。

信息检索常常使用向量空间模型（VSM）来计算语句之间的相关度。通过实验测试一些实例，发现该方法在这里效果并不十分明显。该方法只有当句子所包含的词比较多时效果才好，这是因为它是一种统计的方法，只有当句子包含的词数越多，相关的词才会重复出现，这种统计方法的效果才能体现出来。而在我们的系统中，所面对的是单个的句子，句子中包含的词的个数往往不足以体现这种方法的效果。向量空间模型没有充分利用问题句与回答句中的其他有用信息。例如，关键词之间的距离以及问题句和回答句的长度等信息。这些信息对于抽取回答句有着重要的影响。所以我们提出一种利用语句相关模型来计算问题句和回答句之间相关度的方法。

1）词形相关度

我们知道，语句之间的相关度与语句之间的相同关键词的个数密切相关，语句之间相同非重复关键词个数越多，则其间的相关度也就越高。语句 $Q$、$A$ 的词形相关度为：

$$\mathrm{WordSim}(Q, A)=2\times\frac{\mathrm{SameWC}(Q, A)}{\mathrm{len}(A)+\mathrm{len}(Q)} \tag{1}$$

其中 $\mathrm{len}(Q)$、$\mathrm{len}(A)$ 分别表示语句 $Q$、$A$ 的长度（即 $Q$、$A$ 中所含词的个数），$\mathrm{SameWC}(Q, A)$ 表示 $Q$ 和 $A$ 中相同单词的个数，若一个单词在 $Q$、$A$ 中出现的次数不同，则以出现次数少的计数。

容易证明 $0\leqslant\mathrm{WordSim}(Q, A)\leqslant 1$。

2）句长相关度

我们知道，两个语句的长度越相近，则其间的相关度也就越高，语句 $Q$、$A$ 的句长相关度为：

$$\mathrm{lenSim}(Q, A)=1-\frac{\mathrm{abs}(\mathrm{len}(Q)-\mathrm{len}(A))}{\mathrm{len}(Q)+\mathrm{len}(A)} \tag{2}$$

其中 abs() 表示绝对值。

容易证明 $0\leqslant\mathrm{LenSim}(Q, A)\leqslant 1$。

3）距离相关度

距离相关度是指语句之间关键词距离的大小，两句的相同关键词之间的距离越小，则两句也就越相关；如果同一关键词在句子中出现多次则以产生最小距离的关键词为准。语句 $Q$、$A$ 的距离相关度为：

$$\mathrm{DisSim}(Q, A)=1-\frac{\mathrm{Dis}(A)}{\mathrm{Dis}(A)+\mathrm{Dis}(Q)} \tag{3}$$

其中 $\mathrm{Dis}(Q)$ 表示 $Q$ 中非重复关键词中最左及最右关键词之间的距离，$\mathrm{Dis}(A)$ 表示 $A$ 中与 $Q$ 相同的最左及最右关键词之间的距离。

容易证明 $0\leqslant\mathrm{DisSim}(Q, A)\leqslant 1$

综合考虑上面几个语句间相关度因素，得出问题句和回答句之间总的相关度为：

$$\mathrm{Sim}(Q, A)=\lambda_1\mathrm{WordSim}(Q, A)+\lambda_2\mathrm{LenSim}(Q, A)+\lambda_3\mathrm{DisSim}(Q, A) \tag{4}$$

其中 $\lambda_1$、$\lambda_2$、$\lambda_3$ 表示语句间词形相关度、句长相关度、距离相关度在语句相关度中所占的权重，且 $\lambda_1+\lambda_2+\lambda_3=1$，显然。$0\leqslant\mathrm{Sim}(Q, A)\leqslant 1$

如果问题段落中只有一个问题句（根据统计，这是对话

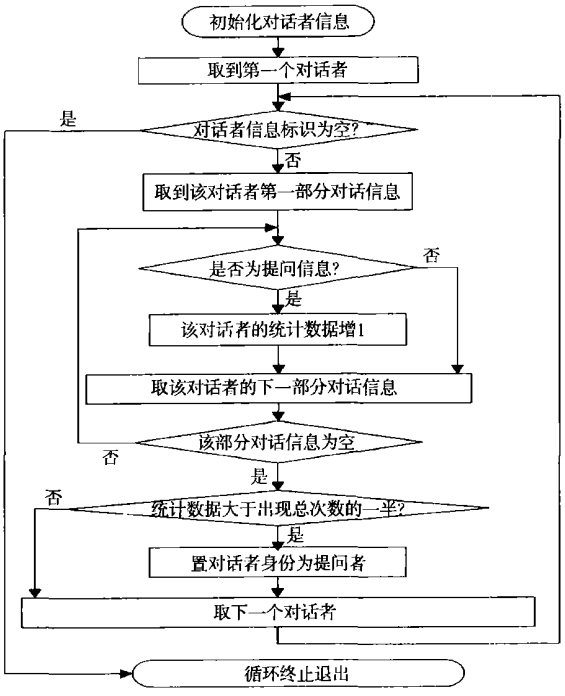


图 **2**　对话者身份判断流程图

式体裁文本中出现最多的形式), 则回答段落中的所有回答句都与问题句关联; 如果问题段落中有多个问题句, 则根据以上语句相关度公式计算出回答段落中每个回答句与问题段落里每个问题句之间的的相关度, 并将回答句关联于相关度最大的问题句。

3.2.4　摘要生成

这个部分中要经过句子的抽取、去重、修饰、无损压缩和摘要的平滑处理等操作, 生成真正意义上的摘要。

在前面得到问答对的基础上, 我们首先计算出每个问答对的权值, 然后从权值较高的问答对开始, 按照启发式规则, 从中抽取合适的句子有机的组成摘要。这样, 当一个问题句被选择后, 其相相应的相关度最大的回答句也应该被选择, 反之亦然。这样就保证了摘要的局部一致性, 从而提高了摘要的全局连贯性。

## 4　实验与分析

我们从人民网、新华网、新浪网等网站上专门搜集了9562篇对话式体裁的中文文本以及 52755 篇非对话中文文本(全部都是 Web 文档)作为语料库, 下面对这个语料库进行体裁判断和问答对识别的分析。

1) 体裁判断

由于在实验中我们对52755篇非对话文本进行体裁判断时, 其错误判断率等于 0, 所以下面我们只给出对话文本的分析结果。

表 **2**　体裁判断平均错分率

| 样本数 | 1000 | 2000 | 3000 | 4000 | 5000 |
|---|---|---|---|---|---|
| 平均错分率(‰) | 4.2 | 3.7 | 3.5 | 3.8 | 3.8 |

表 **2**　体裁判断平均错分率(续)

| 样本数 | 6000 | 7000 | 8000 | 9000 |
|---|---|---|---|---|
| 平均错分率(‰) | 3.6 | 4.0 | 3.8 | 3.8 |

在这里, 我们分别对 1000、2000、……、9000 篇九个级别进行体裁判断错误率的计算, 每个级别均从语料库中抽取 10 次, 且每次均不放回地随机抽取相应级别的篇数, 然后计算出每个级别的最小错分率、最大错分率和平均错分率(如图 3 所示), 其中平均错分率如表 2 所示。从图中可以看出: 随着样本数的增加, 最小错分率和最大错分率均逐渐逼近平均错分率; 样本的平均错分率在某一个数值间浮动, 并有逼近该数值的倾向。

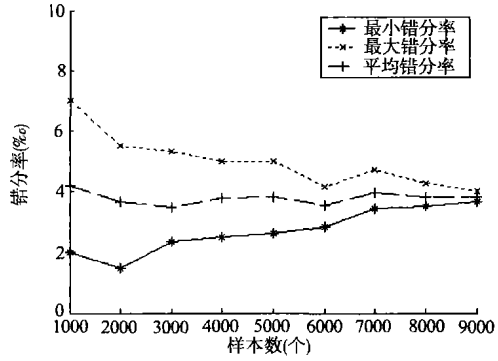


图 **3**　体裁判断错分率

经过初步计算, 9562 篇中约有 36 篇左右的样本体裁判断错误。对于这些文本, 我们经过人工分析, 发现其错误原因主要有两个: 一个是文本长度太短; 另一个是格式不够规范(如表 1 中的其他所示)。总体实验结果表明, 该体裁判断方法是可行有效的。

2) 问答对识别

我们随机抽取 2000 篇系统判断正确的对话文本(样本格式多种多样, 包含答记者问、人物访谈以及网上直播等多种形式), 首先用预处理软件将它们分别规整为纯文本格式, 再请上海交通大学计算机科学与工程系 2002 级 10 名硕士研究生对这些文本进行问题句及其相应回答句的标注工作, 每篇文章里的问题句用[ Q1] , [ Q2] , ……, [ Qn] 进行标识, 与每个问题相对应的回答句用[ A1] , [ A2] , ……, [ An] 进行标识。如:

……

问: [ Q1] 请问辛格外长访华期间是否与中方讨论了地区形势和印巴关系, 中方的立场是什么?

答:[A1] 印度外长辛格访问期间,……交换了看法。[A1] 关于南亚局势特别是印巴关系,……。

问:[Q2] 你对当前巴以局势有何评价?[Q3] 中方在缓和局势方面做了哪些工作?

答:[A2] 中方对最近中东局势急剧恶化深表关切和忧虑。[A2] ……。[A2] 同时,我们也谴责发生在以色列境内的针对无辜平民的暴力活动,呼吁……。

[A3] 中国为缓和当前紧张局势作出了自己的努力和贡献。[A3] ……。[A3] 中方将继续同国际社会一道为缓解目前中东地区的紧张局势作出自己的努力和贡献。

……

在这里,我们取 10 组 $\lambda_1$、$\lambda_2$、$\lambda_3$ 的值进行实验,如果出现如下情况,均属于判断错误:

1)标注中为 Q,判断为 A;

2)标注中为 A,判断为 Q;

3)标注中为[Ai],判断为[Aj] ($i \neq j$);

4)非标注语句,判断为标注语句;

5)标注语句,判断为非标注语句。

我们对人工标注的语料库进行统计,得出语句总数为 43146 个,用我们的系统对这些样本进行问答对的识别,其实验结果如表 3 所示。

表 3 实验结果如

| 组别 | 参数 | | | 召回率 (%) | 正确率 (%) |
|---|---|---|---|---|---|
| | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | | |
| 1 | 1.0 | 0 | 0 | 98.1 | 62.8 |
| 2 | 0.8 | 0.2 | 0 | 91.7 | 70.2 |
| 3 | 0.8 | 0.1 | 0.1 | 92.0 | 74.1 |
| 4 | 0.8 | 0 | 0.2 | 91.5 | 69.7 |
| 5 | 0.7 | 0.2 | 0.1 | 94.2 | 78.9 |
| 6 | 0.7 | 0.1 | 0.2 | 93.3 | 78.4 |
| 7 | 0.7 | 0 | 0.3 | 90.9 | 69.2 |
| 8 | 0.6 | 0.3 | 0.1 | 95.3 | 80.5 |
| 9 | 0.6 | 0.2 | 0.2 | 94.7 | 79.5 |
| 10 | 0.6 | 0.1 | 0.3 | 94.8 | 81.1 |

其中

$$\text{召回率}=\frac{\text{系统标注为问题答案正确的句子数}}{\text{标注为问题答案的句子总数}}\times 100$$

$$\text{正确率}=\frac{\text{标注正确句子数}}{\text{语句总数}}\times 100$$

从实验结果可以看出,当只考虑词形相关度时系统的召回率和正确率都比较低(组 1),如果考虑句子长度的影响,则召回率和正确率有了较大幅度的提高(组 2),如果再考虑距离的影响,则系统的召回率和正确率有了进一步的提高(组 3—10)。

问答对的自动识别与关联方法获得了较高的召回率和识别准确率,这一准确率已经基本满足了对文摘结果连贯性的要求。我们对于标注出错的问答对进行深入分析发现,出错的主要原因有三个个:

1)回答段落与问题段落配对规则过于简单。由于有少数文本在回答段落的后面具有一些叙述性文字,而这些文字与问题段落的任何问题均没有关系,从而使得系统判断错误。

2)样本中存在这样的提问者,他们在全文中总的出现次数很少(一般小于三次),而且他们的对话信息也不具备问题段落的特征,所以用我们的方法来判断,就会将其错误的判断为回答者,进而将其对话信息标注为回答段落。在这种情况下,系统标注问答对的时候,就会将本来的问题段落标注为回答段落,造成错误。

3)由于语句相关度的计算方法本身不能取得很高的准确率,使得系统在问题句和相应回答句的配对的准确率也会受到相应的限制。

后面我们将会针对这三个方面问题进行更加深入的研究,以便进一步提高识别准确率。

## 5 结论

本文主要介绍了对话文本自动摘要系统的若干关键技术和相应的实现技术,其中的对话信息单元(问答对)的自动识别与关联方法,对于对话文本的摘要质量改善起到关键性作用,使得系统在不损失摘要信息量且兼顾摘要的局部一致性的同时,大大提高了摘要的整体连贯性。

参考文献:


[1] 姚天顺. 自然语言理解[M]. 北京:清华大学出版社,1995.

[2] 吴立德. 大规模中文文本处理[M]. 上海:复旦大学出版社,1997.

[3] 徐慧敏. 中文文献自动摘要系统的设计与开发[D]. 上海交通大学硕士论文,1998.

[4] Klaus Zechner. Automatic Summarization of Spoken Dialogs in Unrestricted Domains[C]. Carnegie Mellon University. November 2001.

[5] 崔桓,蔡东风,苗雪雷. 基于网络的中文问答系统及信息抽取算法研究[J]. 中文信息学报,2004,18(3):24—31.

[6] Alon Lavie, Klaus Zechner. Increasing the Coherence of Spoken Dialogue Summaries by Cross—Speaker Information Linking[C]. Carnegie Mellon University. September 2000. ersity.



[作者简介]

陈卫平(1980.4—),女(汉族),山东人,硕士研究生,研究方向:网络信息自动摘要;

王永成(1939.10—),男(汉族),扬州人,教授,博士生导师,研究方向:网络信息智能处理;

刘传汉(1970.3),男(汉族),安徽人,博士研究生,研究方向:智能信息处理。